Extracting Causal Rules from Spatio-temporal Data

Antony Galton[1], Matt Duckham[2] and Alan Both[2]

[1] University of Exeter, UK
[2] RMIT University, Australia

Abstract. This paper is concerned with the problem of detecting causality in spatiotemporal data. In contrast to most previous work on causality, we adopt a logical rather than a probabilistic approach. By defining the logical form of the desired causal rules, the algorithm developed in this paper searches for instances of rules of that form that explain as fully as possible the observations found in a data set. Experiments with synthetic data, where the underlying causal rules are known, show that in many cases the algorithm is able to retrieve close approximations to the rules that generated the data. However, experiments with real data concerning the movement of fish in a large Australian river system reveal significant practical limitations, primarily as a consequence of the coarse granularity of such movement data. In response, instead of focusing on strict causation (where an environmental event initiates a movement event), further experiments focused on perpetuation (where environmental conditions are the drivers of ongoing processes of movement). After retasking to search for a different logical form of rules compatible with perpetuation, our algorithm was able to identify perpetuation rules that explain a significant proportion of the fish movements. For example, approximately one fifth of the detected long-range movements of fish over a period of six years were accounted for by 26 rules taking account of variations in water-level alone.

1 Introduction

In this paper we address the problem of detecting causality in spatiotemporal data. Broadly speaking one might approach this problem in two different ways, which may be labeled *probabilistic* and *logical*. In a probabilistic approach, such as is exemplified by a substantial body of deep and detailed work associated particularly with researchers such as Pearl [10] and Spirtes *et al.* [11], one looks for patterns of conditional dependence and independence in the data which exhibit the characteristic "signatures" of genuinely causal correlations. A typical outcome from this kind of approach is a list of functional dependencies between the values of observational variables, resulting in statements to the effect that one variable has a specific causal influence on another. These approaches may be described as *data-driven* and in particular are appropriate if one has no prior expectation of the form taken by causal laws.

A logic-based approach, in contrast, is driven by a prior conception of the form that causal laws might take, and the process of inferring laws from data is targeted to the discovery of laws of that form. Experiments performed in accordance with this conception may be thought of as investigating the extent to which the given data can be described in terms of laws of a specified form, as opposed to simply trying to discover general causal connections within the data.

We do not here argue for the merits of either approach over the other; the work reported here takes the logical approach rather than the probabilistic, and may be regarded as an investigation into the feasibility of the former.

Specifically, this paper explores the definition of a logical approach to causal analysis of data, capable of generating causal rules of specific logical forms. In §2 we introduce causation and perpetuation in the context of the past literature. In §3 we develop the foundations of logical detection of causal rules, leading to the construction of an algorithm for identifying causal rules from data. The performance of this algorithm is explored first with synthetic data (§4), and subsequently with real data about the environmental context for the movements of fish in the Murray River, Australia (§5). A statistical analysis of these resulting causal rules in §6 demonstrates that the causal rules generated do indeed have explanatory structure in several key respects. Finally, §7 concludes the paper with a look back at the implications for causation in geographic space.

2 Events, Processes, and Causality

The problem with causality, as highlighted by the philosopher David Hume in the 18th century, is that causation is experientially indistinguishable from correlation; that is, causal relations are not themselves overtly present in data but are manifested through correlations which are. But correlations in data can also arise by chance, unconnected with any of the causal mechanisms that in reality gave rise to the data, leading to the appearance of causal dependencies where none are in fact present. And even when a correlation is not the result of chance, it does not mean that there must be a direct causal connection between the correlates, which may instead be independently caused by some common unobserved third element. Thus the problem for anyone seeking to detect causal relations through the analysis of data is how to separate out the genuine cases of causality from such non-causal correlations.

This problem is particularly acute if one takes the kind of "broad brush" view of causality common to many probabilistic approaches. Probabilistic approaches regard causation, appropriately enough, as a relation between events, but then confuse matters by regarding an "event" as anything one can assign a probability to. This is in stark contrast to our normal understanding of events as things that *happen*, i.e., discrete changes in the world. Events in this sense play a central role in causality, but it is important to recognise other forms of causal relation involving processes or states, which differ from events in the manner in which they occupy time [2, 9]. The importance of these temporal distinctions for causal analysis was pointed out in [12, 13].

In the work reported here we take a more focused view of the logical structure of causal relations, which is sensitive to the aspectual distinctions between states, processes, and events in a way that probabilistic approaches typically are not. The ontological framework of our research is taken from [7]; here we recapitulate the main features of this approach.

We take the view that in its strictest sense the verb 'cause' should be understood as naming a relation between discrete events: that is, one event (such as a ball hitting a window) may be said to cause another (such as the window breaking). Loosely, we may also speak of one *process* causing another (e.g., the action of the tides causing erosion) but properly considered this is a different relation because it does not happen at a discrete moment but continues, in an ongoing, cumulative fashion, over a period. For this reason we prefer the term 'perpetuation' for this: the action of the tides perpetuates the erosion process.

Relations of causation and perpetuation apply to individual events or processes: a particular ball impact, at a definite location and time, causes a particular window breaking; the action of the tides along a particular stretch of coastline over a particular period perpetuates the process of erosion along that stretch during that period. But our understanding of the world expects such individual instances to reflect general *laws* referring not to individual occurrent tokens but to *types*. It is, however, generally impossible to formulate a valid law of the form 'Any event of type E_1 causes an event of type E_2', since the causation of an E_2 instance by an E_1 instance is typically dependent on some appropriate enabling conditions: for example, if I turn the door handle and push on the door, the door will open — but only if it is unlocked. The importance of such conditions in causality was emphasised by [1], following [4], and indeed has been widely recognised in the philosophical literature on causality (cf., [5]).

In the light of this, we prefer to formulate *conditional* causal rules along the lines of 'If events of types $E_1, E_2, E_3, \ldots$ occur and conditions $C_1, C_2, \ldots$ hold, then an event of type F will occur'. The conditions can in general be modelled as *states* which either hold or not at each time; but such states may record the state of a process variable, e.g., given a process of variation in water temperature, an example of a state might be that the water temperature exceeds 15°C. These are the kinds of conditions we use in application examples below.

The distinctive feature of the work reported in this paper is that we are looking not just for patterns in data that might betray the existence of causal relationships, but patterns that arise from causal laws of specific forms. Our search for correlations is thus guided by the forms of laws that we hope to find. In the next section we describe the algorithm we use for this.

3 An algorithm for extracting causal rules from data

3.1 The Data

The algorithm takes as inputs one or more *history files*. A history file records the occurrences of events and the values of process variables at every time-step over

some period $T = [0, t_{max}]$, where time-steps are represented as non-negative integers. Events and processes are collectively called *occurrents*: so the set of occurrents, $\mathcal{O}$, may be written as $\mathcal{O} = \mathcal{E} \cup \mathcal{P}$, where $\mathcal{E}$ and $\mathcal{P}$ are the sets of events and processes respectively. Note that $\mathcal{E} \cap \mathcal{P} = \emptyset$.

An event (properly an *event-type*) is represented for the purposes of the algorithm as a function from time-steps to non-negative integers, i.e., for $e \in \mathcal{E}$ we have $e : T \to \mathbb{Z}^+ \cup \{0\}$, where $e(t)$ is the number of distinct occurrences (i.e., *event-tokens*) of e at time t. In any actual application, of course, e will have some semantics specifying its meaning in relation to the application domain; but any such semantics is unknown, and irrelevant, to the algorithm. For many applications the events of interest will be such that $e(t)$ is always either 0 or 1, but this is not invariably the case, and in particular it is not the case for the domain of animal movement we describe later.

Similarly, a process is represented as a function from time-steps to real numbers, so that for $p \in \mathcal{P}$ we have $p : T \to \mathbb{R}$. Typically, though not invariably, they are the discrete-time analogues of continuously-varying real-world functions — for example a process might record the variation in water temperature or water level at a particular station along a river.

3.2 Causal Rules

Amongst events, we regard some as possible *causes* and others as *effects*. These refer to the roles they play in *causal rules*. The most general form of causal rule we handle is:

$$R : [\mathsf{Causes}_R \mid \mathsf{Conditions}_R] \Rightarrow \mathsf{effect}_R \ \mathsf{after}\ \mathsf{Delay}_R,$$

where

- $\mathsf{Causes}_R \subset \mathcal{E}$ is a set of events;
- $\mathsf{Conditions}_R$ is a set of *conditions*, where each condition c is an expression of the form '$v_c^- \leq p \leq v_c^+$', where $v_c^-, v_c^+ \in \mathbb{R}$ and $p \in \mathcal{P}$;
- $\mathsf{effect}_R \in \mathcal{E} \setminus \mathsf{Causes}_R$ is an event distinct from any of the causes;
- Delay_R is a *delay interval* $[d_R^-, d_R^+]$, where d_R^-, d_R^+ are integers such that $0 \leq d_R^- \leq d_R^+$.

In a condition, v_c^- and v_c^+ are the limits of a range within which the value of p_c must fall to satisfy it.

Inclusion of a delay interval does not mean that we are contemplating some mysterious "action at a distance" across time, but simply that the transmission of causal power from cause to effect is mediated by some process that is initiated by the cause and culminates in the effect — e.g., at a traffic intersection, I press the button for the pedestrian signal, and some seconds later (or minutes if I am unlucky) the lights change to enable me to cross.

The causal rule R is *activated* at time t if and only if both:

1. For every $e \in \mathsf{Causes}_R$, $e(t) > 0$.

2. For every $c \in \mathsf{Conditions}_R$, $v_c^- \leq p_c(t) \leq v_c^+$.

An activation of the rule at time t is *explanatory* if the effect predicted by the rule does indeed occur, i.e.:

– For some $d \in \mathsf{Delay}_R$, $\mathsf{effect}_R(t+d) > 0$.

Conversely, an occurrence of effect_R at time t is *explained* by rule R if some activation of R is made explanatory by that occurrence of the effect, i.e.,

– For some $d \in \mathsf{Delay}_R$, R is activated at $t - d$.

It is possible for a rule-activation to explain more than one occurrence of its effect, and also for an effect to be explained by more than one rule-activation. These may be regarded as unsatisfactory situations from the perspective of some real-world applications, but in others may be perfectly acceptable, e.g., one and the same environmental event may trigger migration in many individual fish; and migration by a fish may be triggered by two different environmental events each of which would be sufficient on its own to cause it.

From the general form of rule as presented here, a number of special cases can be identified that are of interest. If $\mathsf{Conditions} = \emptyset$, we have an *unconditional* rule, which can be written in simplified form as

$$R : \mathsf{Causes}_R \Rightarrow \mathsf{effect}_R \ \ \mathsf{after}\ \mathsf{Delay}_R.$$

If $\mathsf{Delay}_R = [d, d]$ we have a *one-delay* rule, which can be written as

$$R : [\mathsf{Causes}_R \mid \mathsf{Conditions}_R] \Rightarrow \mathsf{effect}_R \ \ \mathsf{after}\ d$$

and in the special case $d = 0$ we have a *simultaneous causation* rule, written

$$R : [\mathsf{Causes}_R \mid \mathsf{Conditions}_R] \Rightarrow \mathsf{effect}_R.$$

For any of these rules it will sometimes be convenient to abbreviate the part before the $\Rightarrow$ as $\mathsf{antecedent}_R$, which is neutral as to its composition out of causes and conditions, e.g.,

$$R : \mathsf{antecedent}_R \Rightarrow \mathsf{effect}_R \ \ \mathsf{after}\ \mathsf{Delay}_R.$$

3.3 The Problem

The problem which the algorithm is designed to solve may be stated simply as follows: Given a data set in the form described in §3.1, we seek a set of rules $\mathcal{R}$ which, as nearly as possible, accounts fully for the data, in the following sense:

1. For each $t \in T$ and $R \in \mathcal{R}$, if R is activated at t then it is explanatory, i.e., effect_R occurs after an admissible delay.
2. For each occurrence of each effect f in the data, there is a rule $R \in \mathcal{R}$ which explains it, i.e., $f = \mathsf{effect}_R$ and R is activated within an admissible delay time preceding the occurrence.

These rules can be roughly characterised as "no false positives" and "no false negatives" respectively, though the precise interpretation of these terms in the present context is delicate and will be discussed further below.

With real-world data it is unrealistic to expect to find a rule-set which fully accounts for the data in this sense, which is why we add the caveat 'as nearly as possible' to the problem statement. To interpret this we need to find a measure of *how nearly* a rule-set fully accounts for the data. This is discussed in §3.4.

3.4 Evaluating a rule set

Given some data and a set of rules (however these have been discovered, whether by the algorithm described here or in some other way), we need a principled way of evaluating the rules with respect to the data. For this purpose two commonly used measures are *precision* and *sensitivity*. In general, for a rule of the form 'If P then Q' these are defined as

$$\mathit{precision} = \frac{TP}{TP + FP} \qquad \mathit{sensitivity} = \frac{TP}{TP + FN}$$

where

- TP is the number of *true positives*, i.e., instances satisfying both P and Q,
- FP is the number of *false positives*, i.e., instances satisfying P but not Q,
- FN is the number of *false negatives*, i.e., instances satisfying Q but not P.

Our problem is how to define these quantities for a causal rule of the form

$$R : \mathsf{antecedent}_R \Rightarrow \mathsf{effect}_R \ \ \mathsf{after}\ [d^-, d^+].$$

In particular, what do we mean by an 'instance'?

In the case of TP, we could take either a *cause-centred* (cTP) or an *effect-centred* (eTP) approach as follows:

- cTP is the number of explanatory activations of R
- eTP is the number of occurrences of effect_R which are explained by R.

In general, these figures will be different. For the other two quantities of interest, it is natural to count FP in the cause-centred way, and FN in the effect-centred way, as follows:

- cFP is the number of non-explanatory activations of R
- eFN is the number of occurrences of effect_R that are not explained by R

We now define *cause-centred precision* and *effect-centred sensitivity* as follows:

$$\textit{c-precision} = \frac{cTP}{cTP + cFP} \qquad \textit{e-sensitivity} = \frac{eTP}{eTP + eFN}$$

Thus *c-precision* measures what fraction of the rule activations are explanatory, and *e-sensitivity* measures what fraction of occurrences of the effect are explained by the rule.

These definitions can be used to evaluate an individual rule; to evaluate a set of rules $\mathcal{R}$ for the same effect we use

– cTP: the number of explanatory activations of a rule in $\mathcal{R}$
– cFP: the number of non-explanatory activations of a rule in $\mathcal{R}$
– eTP: the number of occurrences of effect explained by at least one rule in $\mathcal{R}$
– eTP: the number of occurrences of effect not explained by any rule in $\mathcal{R}$

The harmonic mean of the *c-precision* and *e-sensitivity* is called the F_1 score and provides a useful single measure against which rules can be ranked:

$$\mathrm{F}_1 = 2\Big(\frac{\textit{c-precision} \cdot \textit{e-sensitivity}}{\textit{c-precision} + \textit{e-sensitivity}}\Big).$$

3.5 The algorithm

The algorithm is presented below as Algorithm 1. Here we give an informal explanation of it to help the reader understand how it works, as well as some pertinent observations. Note that, in the algorithm, we use $\mathcal{F}$ to refer to the set of effects to be explained.

The algorithm is guided in its search for causal rules by the strict form to which any such rule must adhere. For each effect f, and each subset E of the events available to act as causes, we consider whether any of the data for f can be explained by a rule whose cause-set is E.[3] Such a rule could only be activated at those times T_E at which every event in E occurs; we can therefore immediately discard any set E for which there are no such times, along with any supersets of that set (line 5). If on the other hand T_E is non-empty, we need to consider whether each time in T_E is followed by an occurrence of f within $d_{\max}$ time-steps, where $d_{\max}$ is the maximum allowed delay for a rule (set by the user).

Let D_T be the set of all delays d in the range $[0, d_{\max}]$ for which some time in T_E is followed by an occurrence of f after a delay of d time steps. Any causal rule generating some of these occurrences of f from E must have a delay interval encompassing some of the delays in D_T. Hence if D_T is empty we can discard E and all its supersets (line 10).

If E is still not discarded, then we have a set of times T_E at which all the putative causes in E occur, and for each of these times there may or may not be an occurrence of f within a delay in the set D_T. The times for which such an occurrence exists are put in the set T^+, the rest in T^- (line 12).

If T^- is empty, this means that *whenever* all of E occur, f occurs after a suitable delay. Letting d^- and d^+ be the minimum and maximum delays in D_T, we can set up the unconditional rule '$E \Rightarrow f$ after $[d^-, d^+]$' (line 15), and this is guaranteed to generate no false positives for the data, i.e., to satisfy the first condition in §3.3. (There may of course be false negatives since there may be more than one rule for effect f, with different cause-sets.)

[3] At line 3 of the algorithm we are required to iterate over the power set of $\mathcal{E}$. Since this leads to combinatorial explosion if $\mathcal{E}$ is too big, we in practice restrict the iteration to subsets of $\mathcal{E}$ up to some predetermined size. In any case we are most likely to be interested in rules with a small number of causes in the antecedent.

Algorithm 1: The rule-detection algorithm

```
1  Let R = ∅;
2  foreach f ∈ F do
3      foreach E ⊆ E do
4          Let T_E be the set of t ∈ T such that e(t) ≥ 1 for every e ∈ E.;
5          if T_E = ∅ then jettison E and all its supersets;
6          else
7              foreach t ∈ T_E do
8                  let D_t be the set of d ∈ [0, d_max] such that f(t + d) = 1;
9              Let D_T = ⋃_{t∈T} D_t;
10             if D_T = ∅ then jettison E and all its supersets;
11             else
12                 Let T⁺ = {t ∈ T_E | D_t ≠ ∅} and T⁻ = {t ∈ T_E | D_t = ∅};
13                 if T⁻ = ∅ then **we have an unconditional rule**
14                     Let d⁻ = min(D_T) and d⁺ = max(D_T);
15                     R ← R ∪ {[E | ∅] ⇒ f after [d⁻, d⁺]};
16                 else **we look for conditional rules**
17                     foreach p ∈ P do
18                         Sort T_E w.r.t. the value of p at each time. Call the
                           sorted list T_E^s;
19                         Create a new list T_E^w from T_E^s such that the ith element
                           of T_E^w is the number of elements of T⁺ occurring in the
                           subsequence of T_E^s with indices in the range [i − h, i + h]
                           (where h is a pre-determined constant);
20                         Now find all maximal subsequences of T_E^w of length
                           greater than 2h + 1 in which the values are all positive;
21                         foreach subsequence covering indices i_1, ..., i_n do
22                             Let t_0 and t_1 be the i_1th and i_nth elements of T_E^s
                               and put v⁻ = p(t_0) and v⁺ = p(t_1);
23                             Let D_T^p = ⋃{D_t | p(t) ∈ [v⁻, v⁺]} and let
                               d⁻ = min(D_T^p) and d⁺ = max(D_T^p);
24                             R ← R ∪ {[E | v⁻ ≤ p ≤ v⁺] ⇒ f after [d⁻, d⁺]};
25 Remove from R any rule that is covered (see below) by another rule in R;
```

If on the other hand T^- is not empty, then not all occurrences of E are followed by f within the acceptable delay time. In this case, we might still find an unconditional rule that admits exceptions (false positives), and if we are interested in these we can relax the condition $T^- = \emptyset$ at line 13. But with this condition in place, we must proceed to the search for conditional rules (lines 16–24). To this end we consider in turn each of the processes available to supply conditions (remember that a condition takes the form $v^- \leq p \leq v^-$, where $[v^-, v^+]$ is the range within which the process variable p must fall for the condition to be satisfied).

Suppose that in fact all the data for f could be accounted for by a single rule '$[E \mid v^- \leq p \leq v^+] \Rightarrow f$ after $[d^-, d^+]$'. Since $T^- \neq \emptyset$, there are occurrences of E that are not followed by f within an appropriate delay. The non-occurrence

of f must be explained by the value of p being outside the range $[v^-, v^+]$ at that time. If, therefore, we sort the times in T_E with respect to the value taken by p at those times to give the sequence T_E^s (line 18), marking each time "good" or "bad" according as the effect f does or does not occur then, the "good" times will form a consecutive run within the sequence, with the values of p at the start and end points of this run bracketed by the "true" values v^- and v^+ — and this is the maximal run of consecutive elements within the sequence for which this is the case. If just one rule fully accounts for the data, a close approximation to it (differing only in the precise value-range in the condition) can be discovered by the above procedure.

In general, however, we expect there to be other rules, with the same effect, whose presence prevents the simple procedure above from working, it being unlikely that the "good" times in the value-sorted sequence will form a single consecutive run. In this case, two immediate remedies suggest themselves:

- On the one hand, we could simply take as our v^- and v^+ the smallest and greatest values attained by p on T^+. The resulting rule is guaranteed to exclude any false negatives, since every actual occurrence of f within the delay range is covered, but it may admit many false positives (the gaps in the sequence of "good" points).
- On the other hand, since there is no single run of consecutive "good" values, we could look for *all* such runs in the sequence and construct a new rule for each, using the extreme p values within that run as our v^- and v^+ for that rule. This method will create a set of rules for f which are guaranteed to exclude false positives (since none of the rules will be activated at any of the "bad" points) but at the cost of a proliferation of rules each of which allows many false negatives.

The method actually used in the algorithm is a compromise between these two approaches, and is found in practice to generate rules with fewer false positives than the first and fewer false negatives than the second.

What we do is to run a sliding "window" of length $2h+1$ along the sequence T_E^s (with suitable adjustments for the first and last h positions in the list), recording in T_E^w the total numbers of "good" points within the window at each position (line 19). This achieves a smoothing effect on the sequence, allowing us to identify the ranges of values for p within which the occurrence of the effect is more frequent than elsewhere. These show up as maximal runs of positive values in T_E^w; they are the ranges we use in the conditions for rules (lines 20-24).

Finally, having collected a set of rules for effect f, we discard any which are superfluous because they are covered by other rules in the set (line 25). Rule R_1 *covers* rule R_2 with respect to the data so long as $\mathsf{effect}_{R_1} = \mathsf{effect}_{R_2}$ ($= f$, say), every occurrence of f in the data that is explained by R_2 is also explained by R_1, and every non-explaining activation of R_2 is also an activation of R_1. In this case R_2 is superfluous and can be dropped from the rule-set.

It should be noted that the algorithm, as currently constituted, can only generate rules with $|\mathsf{Conditions}_R| \leq 1$.

4 Working with synthetic data

The algorithm was first tested on synthetic data sets, generated using artificial causal rules. This form of synthetic data set enabled investigation of how well the algorithm could retrieve known rules from the data. For this it was necessary to: (a) define occurrents to feature in the antecedents of the rules; (b) generate histories for those occurrents over an adequate number of time-steps; and then (c) determine the activation history for each rule and thereby generate histories for the effects of the rules. The occurrent histories and effect histories were then used as inputs to the rule-detection algorithm.

Several types of occurrent were defined, as follows:

1. Events:
   - Periodic events, specified by the number of time-steps from one occurrence to the next
   - Random events, specified by the probability of occurrence at any time-step
2. Processes:
   - Sinusoidal processes, specified by the period
   - Gaussian processes, stipulated to have mean 0 and standard deviation 1
   - Markovian processes, in which the differences between the values at consecutive time-steps have a Gaussian distribution with mean 0 and standard deviation 0.1.

4.1 Experiment 1

For this first set of experiments, the occurrents used were those listed in Table 1, and the rules used were those listed in Table 2. It will be noted that some of the occurrents did not feature in any of the rules. This does not mean that they played no role in any of the experiments with this rule-set. They were available to the rule-detection program, which could therefore look for rules featuring these occurrents. Thus these occurrents acted as "red herrings", and indeed it will be seen from the results in Table 3 that in two cases the best rules found did feature occurrents from this set (pGauss2 and pMarkov2).

Three runs were performed with this set of occurrents and rules, each with 1000 time-steps. For each rule found by the algorithm, the *c-precision* and *e-sensitivity* were computed, and from these the F_1 score was derived. For each effect, the rule with highest F_1 score is reported in Table 3. It will be noted that the rule for effect f1 is deterministic, since the antecedent is a periodic event, with the same occurrences in each run, and the delay interval is a single point. This does not mean that the detection program has an identical task for this effect in all three runs, since it does not "know" in advance that it was generated by a rule of that type, and therefore is also looking for rules whose activations may differ between the runs. None the less, both for this effect and the non-deterministic f2 and f3, the program reliably found the correct rule on each occasion; these are, of course, the effects generated by unconditional rules.

Occurrent	Type	Parameters
pSineHigh	sinusoidal process	period 10
pSineMedium	sinusoidal process	period 18
pSineLow	sinusoidal process	period 32
pGauss1	Gaussian process	
pGauss2	Gaussian process	
pMarkov1	Markovian process	
pMarkov2	Markovian process	
ePeriHigh	periodic event	period 9
ePeriMedium	periodic event	period 24
ePeriLow	periodic event	period 50
eRandomHigh	random event	probability 0.4
eRandomMedium	random event	probability 0.25
eRandomLow	random event	probability 0.1

Table 1. Occurrents used in Experiment 1

Rule	*Activations*
ePeriLow $\Rightarrow$ f1 after 5	20, 20, 20
ePeriMedium $\Rightarrow$ f2 after $[3, 6]$	42, 42, 42
ePeriHigh, eRandomHigh $\Rightarrow$ f3 after $[0, 4]$	42, 57, 54
[eRandomHigh $\mid 0 \leq$ pSineMedium ≤ 1] $\Rightarrow$ f4 after 2	200, 238, 226
[eRandomMedium $\mid 0 \leq$ pSineHigh ≤ 1] $\Rightarrow$ f5 after $[0, 3]$	162, 149, 152
[ePeriHigh, eRandomHigh $\mid 0 \leq$ pGauss1] $\Rightarrow$ f6 after 3	21, 26, 27
[ePeriMedium, eRandomMedium $\mid 0 \leq$ pMarkov1] $\Rightarrow$ f7 after $[2, 4]$	4, 7, 9

Table 2. Rules used to generate data for Experiment 1, with the number of times that each was activated in 1000 time steps for runs 1, 2, and 3.

With the conditional rules, the program had a less easy time of it, but still managed to find good approximations to the "true" rules in almost every case. The exceptions were for f7 in runs 1 and 3, where the program identified the correct causes, but mistook the conditions, attributing the effect to conditions involving the processes pMarkov2 and pGauss2 which in fact figure in none of the correct rules. The poor performance for this effect can be explained by the fact that in each run it occurred at less than 1% of the time-steps, so there was not enough relevant data for the rule-detection algorithm to work on, with the result that a spurious non-causal correlation happened to provide a better fit to the data than the best approximation to the true rule discoverable by the algorithm.

In summary, these and other experiments performed with synthetic data demonstrated that the algorithm was, in most cases, able to retrieve from synthetic data the causal rules that generated it. In instances where the algorithm failed, there were frequently plausible explanations for that failure, such as a lack of relevant data generated by a specific rule.

Run	*Best rules found*	*CP,ES*
1	ePeriLow $\Rightarrow$ f1 after 5	100, 100
2	ePeriLow $\Rightarrow$ f1 after 5	100, 100
3	ePeriLow $\Rightarrow$ f1 after 5	100, 100
1	ePeriMedium $\Rightarrow$ f2 after $[3, 6]$	100, 100
2	ePeriMedium $\Rightarrow$ f2 after $[3, 6]$	100, 100
3	ePeriMedium $\Rightarrow$ f2 after $[3, 6]$	100, 100
1	ePeriHigh, eRandomHigh $\Rightarrow$ f3 after $[0, 4]$	100, 100
2	ePeriHigh, eRandomHigh $\Rightarrow$ f3 after $[0, 4]$	100, 100
3	ePeriHigh, eRandomHigh $\Rightarrow$ f3 after $[0, 4]$	100, 100
1	[eRandomHigh $\mid -0.342 \leq$ pSineMedium ≤ 0.985] $\Rightarrow$ f4 after $[0, 4]$	92, 100
2	[eRandomHigh $\mid 0.643 \leq$ pSineMedium ≤ 0.985] $\Rightarrow$ f4 after $[0, 4]$	100, 76
3	[eRandomHigh $\mid -0.342 \leq$ pSineMedium ≤ 0.985] $\Rightarrow$ f4 after $[0, 4]$	97, 100
1	[eRandomMedium $\mid -0.588 \leq$ pSineHigh ≤ 0.951] $\Rightarrow$ f5 after $[1, 5]$	86, 90
2	[eRandomMedium $\mid 0 \leq$ pSineHigh ≤ 0.951] $\Rightarrow$ f5 after $[0, 4]$	100, 100
3	[eRandomMedium $\mid -0.588 \leq$ pSineHigh ≤ 0.951] $\Rightarrow$ f5 after $[2, 6]$	71, 84
1	[ePeriHigh, eRandomHigh $\mid 0.03 \leq$ pGauss1 ≤ 1.959] $\Rightarrow$ f6 after 3	100, 100
2	[ePeriHigh, eRandomHigh $\mid 0.033 \leq$ pGauss1 ≤ 1.96] $\Rightarrow$ f6 after 3	100, 100
3	[ePeriHigh, eRandomHigh $\mid 0.065 \leq$ pGauss1 ≤ 2.975] $\Rightarrow$ f6 after 3	100, 100
1	[ePeriMedium, eRandomMedium $\mid -4.249 \leq$ pMarkov2 ≤ -0.76] $\Rightarrow$ f7 after $[2, 3]$	100, 75
2	[ePeriMedium, eRandomMedium $\mid 0 \leq$ pMarkov1 ≤ 1.761] $\Rightarrow$ f7 after $[2, 4]$	100, 100
3	[ePeriMedium, eRandomMedium $\mid 0.482 \leq$ pGauss2 ≤ 1.543] $\Rightarrow$ f7 after $[2, 4]$	100, 67

Table 3. The best rule discovered by the program for each effect in each run of Experiment 1, with *c-precision* and *e-sensitivity* (expressed as percentages)

5 Working with real data

5.1 Fish movement data set

The real-world data set used for this study was one we had previously worked with, as described in [3]. Lyon and collaborators [8] gathered data on fish movement in the Murray River system in south-eastern Australia. Over 1000 individual fish were tagged with radio transmitters, and their movements were monitored by 18 river-side radio receivers located at strategic positions along the river, which thereby divided the river and its tributaries into 24 zones, labelled *a*–*x* (see Figure 1). The movement of tagged fish between the zones was tracked over a period of six years, during which time a number of environmental variables were also monitored, including water temperature, water level, and salinity. The environmental variables were recorded at a coarser spatial granularity than the fish movements, since the recording stations were more widely spaced along the river than the radio receivers: thus the values for these variables in a zone are taken to be those recorded at the nearest station to the zone.

The data thus consisted of records of the following types:

– For each environmental variable, a record of its value at each recording station on each day of the period of study;

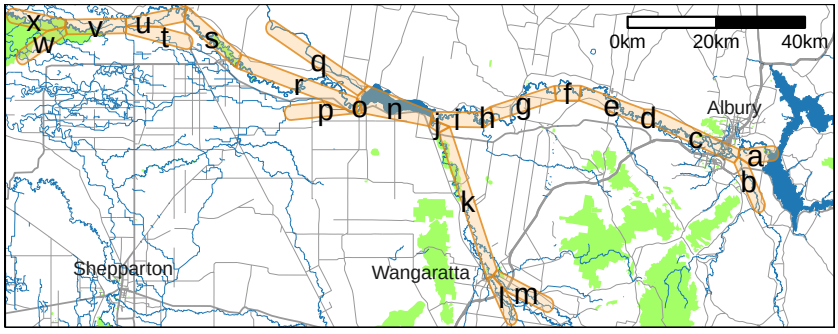


Fig. 1. Map of study area, Murray River, Australia, showing river zones a–x

– A collection of records of zone-boundary crossings by individual fish, where each record takes the form 'fish i moves from zone z_1 to zone z_2 on day d'.

The aim of our study was to determine to what extent the movement of fish was causally influenced by the variations in the environmental variables.

To this end, fish-movement event types were defined as follows. For each pair z_1, z_2 of adjacent zones, where z_2 is downstream from z_1, the event $\mathsf{z_1 \backslash z_2}$ occurs whenever a fish moves from z_1 to z_2, and the event $\mathsf{z_2/z_1}$ occurs whenever a fish moves from z_2 to z_1. Note that it is possible for there to be several occurrences of any one of these events on any given day.

Two sets of experiments were performed using this data, which are reported in the next two sections.

5.2 Experiment 2

For this experiment, we looked for unconditional rules relating fish movement events to a certain set of events defined in terms of the environmental variables. For each environmental variable v and each group of zones G relating to a given recording station for that variable, we defined the event $v\mathsf{3q}(G)$ as occurring whenever the value of v recorded at G crossed from the third to the fourth quartile of its range. Thus for example the event $\mathsf{wl3q(cd)}$ stands proxy for 'onset of high water level in zones c and d'.

The algorithm was asked to look for rules with these quartile-boundary crossing events as causes, and the fish-movement events as effects. We did not pursue this line of enquiry beyond the initial stages as it became clear that the results were somewhat disappointing. Here we present just those results obtained when we looked for rules relating the environmental events $\mathsf{wl3q(cd)}$, $\mathsf{wl3q(efgh)}$, $\mathsf{wl3q(ijklm)}$ and the downstream boundary-crossing events $\mathsf{c\backslash d}$, $\mathsf{d\backslash e}$, $\mathsf{e\backslash f}$, $\mathsf{f\backslash g}$, $\mathsf{g\backslash h}$, $\mathsf{h\backslash i}$, $\mathsf{i\backslash n}$, $\mathsf{k\backslash j}$, $\mathsf{m\backslash k}$. Only 26 rules were found, all with F_1 scores below 30%. The highest ranking rules, with their F_1 scores, are listed in Table 4.

Rule	F_1
wl3q(efgh) ⇒ d\e after $[3, 10]$	0.29
wl3q(cd) ⇒ d\e after $[0.10]$	0.26
wl3q(cd) ⇒ e\f after $[1, 10]$	0.24
wl3q(efgh) ⇒ e\f after $[1, 9]$	0.24
wl3q(ijklm) ⇒ i\n after $[0, 9]$	0.24
wl3q(cd) ⇒ c\d after $[5, 8]$	0.20
wl3q(cd) ⇒ f\g after $[0, 9]$	0.20
wl3q(efgh) ⇒ c\d after $[3, 9]$	0.16
wl3q(efgh) ⇒ f\g after $[0, 10]$	0.16

Table 4. The top-ranking rules by F_1 score from Experiment 2 (part)

On the face of it, some of these rules make more sense from a spatial point of view than others. We would expect the strongest causal influence on a fish's movement between two zones to come from the environmental conditions within the zone from which the fish is moving. Thus of the first two rules in the table, the one relating d\e to wl3q(cd) is *prima facie* more "sensible" than the one relating the same effect to wl3q(efgh). In fact the presence of both rules, with comparable F_1 scores, reflects the high correlation between the values of wl3q(cd) and wl3q(efgh) (correlation coefficient 0.9768). This high correlation explains why each rule with the former event as cause is paired with a rule with the latter event, with similar F_1 score. Equally, the *low* correlations between these two values and wl(ijklm) (-0.031 and -0.047 respectively) account for the absence of similar pairings with rules involving that event.

5.3 Discussion

The disappointingly low F_1 scores found in Experiment 2 prompted us to revise our ideas about the kind of causal rule we should be looking for. The initial idea was that initiation of fish movement should be triggered by some environmental event, in accordance with the principle that events are caused by events, so quartile-boundary crossing was used as a way of deriving candidate events from the processes provided in the data. However, there is something rather arbitrary about this choice of events, and coupled with the fact that the crossing of zone-boundaries is also a rather crude proxy for initiation of fish-movement, it is not surprising that the rules discovered, although not implausible, were rather weak.

On reflection, it seemed that rather than looking for rules relating environmental events to the *initiation* of fish movement, it would be more fruitful to look for rules relating environmental processes to the fish movement, considered as a process itself. The kind of causality considered in our third experiment is thus *perpetuation* rather than causation in the narrow sense, the zone-crossing events now being considered as proxies for upstream or downstream movement *processes*.

6 Experiment 3: Exploring processes and perpetuation

In order to handle perpetuation, we need to specify rules without causes in the antecedent. To model this, we require rules in which Causes is empty, so that all the burden of causality is borne by the conditions. In order to work with this kind of rule using the algorithm, a "dummy" event was generated which occurred at every time-step. This was achieved simply by defining the event (called always) as a random event with probability 1. For clarity, the rule

$$[\mathsf{always} \mid \mathsf{Conditions}] \Rightarrow \mathsf{effect\ \ after\ Delay}$$

will be written in shorter form as

$$\mathsf{Conditions} \Rightarrow \mathsf{effect\ \ after\ Delay}.$$

We shall call these "Always-rules".

This section reports on a systematic exploration of the rules generated by the algorithm when tasked with identifying *perpetuation* in the fish data set. For brevity, we focus on the rules generated from causal analysis of data about water levels and movement. However, the same analysis has been conducted on the water-temperature data with congruent results.

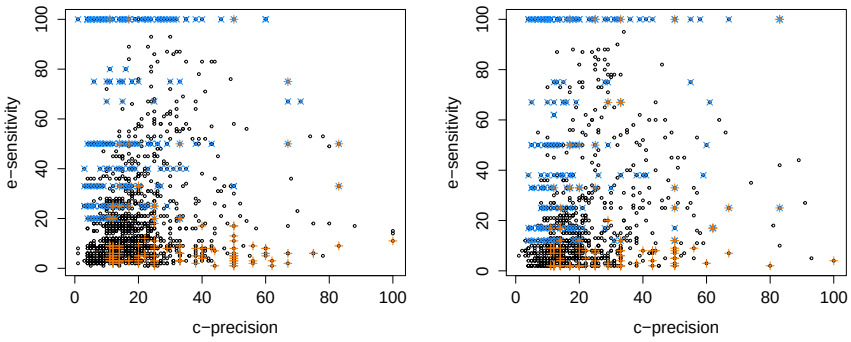
6.1 Support

The algorithm is able to identify a large number of candidate rules from the data set. For example, more than 1000 rules are found for each upstream and downstream movement in response to water level. However, many of these rules are derived from conditions or events that occur only a handful of times. Figure 2 shows a scatterplot of *e-sensitivity* and *c-precision* of rules generated for upstream and downstream movement in response to water levels.

Figure 2 highlights those rules that relate to less than 10 instances of conditions (orange "+") or to less than 10 instances of effects (blue "×"). It is immediately noticeable that rules supported by few condition instances also have lower *e-sensitivity*, and similarly rules supported by few effect instances tend to have lower *c-precision*. Hypothesis testing confirms this visual expectation, significant at the 1% level. Taking this result as a evidence of overfitting, those rules that were supported by less than 10 condition or effect instances in the data were excluded from the subsequent analyses.

6.2 Spatial coincidence

Next we examined the spatial coincidence between conditions and effects. The causal analysis is agnostic about whether an effect is in any way spatially related to its condition. As in Experiment 2 (§5.2), it was found that many of the rules generated relate conditions in one zone to effects in a different zone. However, one might hope that "sensible" rules would relate conditions in one zone to effects in the same zone.



a. Upstream movement

b. Downstream movement

Fig. 2. Scatterplot of rule *e-sensitivity* against *c-precision*, highlighting rules with fewer than 10 instances of conditions (orange "+") and fewer than 10 instances of effects (blue "×").

We tested whether those rules that related conditions to spatially proximal effects (i.e., where the condition was spatially coincident with the start of movement) tended to have higher F_1 scores than rules that related conditions to spatially distal effects. A non-parametric Wilcoxon rank sum hypothesis test indicated that there was no evidence to support the hypothesis that spatially coincident rules have higher F_1 scores ($p = 0.39$ for upstream and $p = 0.88$ for downstream movement, which leads us to fail to reject the null hypothesis that proximal and distal conditions are drawn from the same population of F_1 scores).

Thus, as in §5.2, the data do not support the expectation that F_1 scores are higher for spatially proximal effects; indeed it appears that rules that relate conditions to distal effects are just as likely to have good *c-precision* and *e-sensitivity* as those that relate conditions to proximal effects. As we have already seen, such rules can potentially occur both as an effect of spatial autocorrelation in conditions and as a granularity effect (cf. §5.2). Nevertheless, we restricted our subsequent analyses to examine only "sensible" rules (where condition and effects are spatially proximal) on the grounds that such rules are more meaningful (even if our data did not indicate that they were statistically distinct).

6.3 Shuffled data

In this context, we examined the degree to which the rules might still relate to meaningful patterns, rather than arbitrary overfitting, by repeating the causal analysis with a "shuffled" data set. In our shuffled data set, observations of environmental variables were arbitrarily reassigned to randomly selected zones (e.g., the water level in zone a might be reassigned to zone f at time t_1 and reassigned

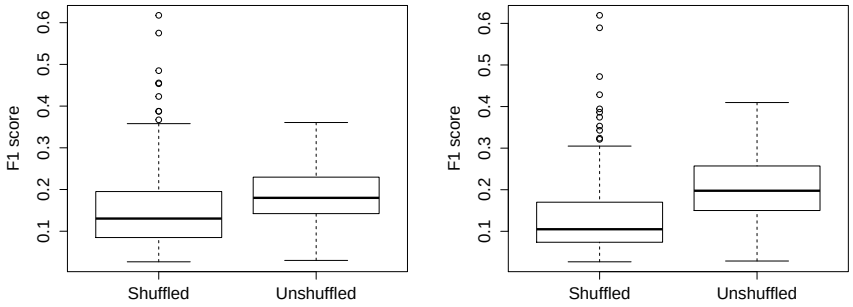
to zone p at time t_2, and so on). This process ensures that any structure in the data resulting from causal relationships is lost, while still allowing comparison with the unshuffled data set (since the movements are unchanged, and the distribution of the total set of environmental variables is unchanged).

There are two main reasons in this case for preferring shuffling to more conventional cross-validation (where the algorithm results are applied to a reserved portion of the data). First, cross-validation is sensitive to how the data set is partitioned. Spatial, seasonal, and longer-term variations (including drought conditions in the earlier years of the study) are expected to lead to statistical non-stationarity in the data. Consequently, by partitioning the data, especially with respect to time or space, we would run the risk that the reserved portion exhibits different properties to the training data. Second, cross validation cannot yield information about the "correct" rules, since (unlike in our experiments with synthetic data) we have no ground truth in the form of causal rules with which to compare the results, such rules being manifested only through correlations in the data (as discussed in §2). Cross-validation will only tell us how sensitive our results are to partitions of the data. This information is already implicitly available in the support for each rule, and indeed rules with low support are discarded anyway (§6.1). By contrast, shuffling allows us to create a second data set for cross-validation that has identical statistical properties (same numbers, timing, and locations of movement events, same numbers and distributions of process variables) to the original, unshuffled data. Any spatial relationships between causes and effects are thus scrambled in the shuffled data set. As a consequence, any rules inferred from the shuffled data are a priori examples of overfitting, and any difference between the results for the unshuffled and shuffled data sets can be ascribed to underlying spatial patterns in the unshuffled data set.

Figure 3 shows the boxplots of F_1 scores for rules generated from both shuffled and unshuffled data sets. In all cases, the F_1 scores for the unshuffled data set are significantly higher (at the 1% level, $p < 0.0001$) than for the shuffled data set. Thus we may infer that the rules generated do indeed derive from *some* meaningful patterns of movement, and are not purely overfitting.

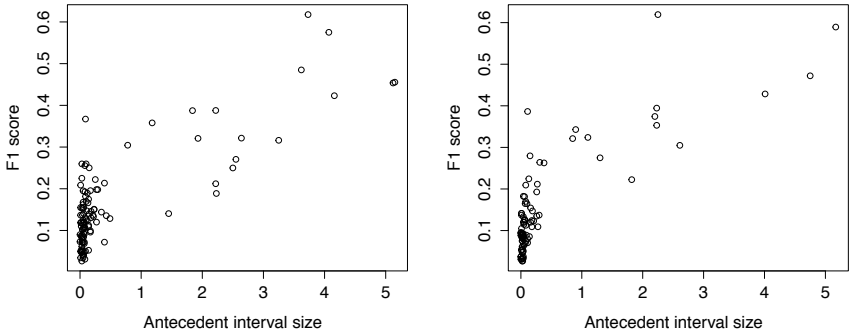
6.4 Condition value ranges

Looking at the rules themselves, it was noticeable that the size of the value range in the antecedent (condition) for a rule was strongly correlated with the F_1 score for that rule (Figure 4). In other words, rules with larger ranges for the environmental variables in the antecedent tended to be associated with larger F_1 scores. This is encouraging as, in general, such rules can be regarded as stronger: they "say more," since they make assertions about a wider range of instances and therefore it takes less to falsify them. (Conversely, rules with larger delay intervals can be regarded as weaker, since they "say less" about precisely what effects are expected to result from a condition.)



a. Upstream movement b. Downstream movement

Fig. 3. Boxplot of F_1 scores for rules generated from shuffled and unshuffled data sets.



a. Upstream movement (Pearson's r=0.807)

b. Downstream movement (Pearson's r=0.813)

Fig. 4. Scatterplot of condition interval length against F_1 score.

6.5 Top-ranked rules

Finally, we looked at the top-ranked rules (in terms of F_1 score) for each effect, in Table 5. In total, these 15 rules accounted for more than 20% of all upstream movements found in the data set. Combined with the 11 top-ranked rules for downstream movement, which accounted for more than 18% of all downstream movements, a total of 26 rules accounted for a significant minority (approximately one-fifth) of all movements. Given that water level is but one potential driver of movement, and that our rules take but one, simple form, this small set

of rules does seem to provide a surprisingly compact representation of almost one fifth of the data set.

Best rule found	F_1 score
$2.79 \leq \mathsf{wl(cd)} \leq 4.72 \Rightarrow \mathsf{d/c}$ after $[0,5]$	0.32
$2.39 \leq \mathsf{wl(efgh)} \leq 5.03 \Rightarrow \mathsf{e/d}$ after $[0,5]$	0.32
$2.81 \leq \mathsf{wl(efgh)} \leq 5.03 \Rightarrow \mathsf{f/e}$ after $[0,5]$	0.38
$1.77 \leq \mathsf{wl(efgh)} \leq 1.92 \Rightarrow \mathsf{g/f}$ after $[0,5]$	0.25
$0.77 \leq \mathsf{wl(efgh)} \leq 1.55 \Rightarrow \mathsf{h/g}$ after $[0,5]$	0.30
$126.41 \leq \mathsf{wl(ijklm)} \leq 131.53 \Rightarrow \mathsf{i/h}$ after $[0,5]$	0.45
$126.85 \leq \mathsf{wl(ijklm)} \leq 128.69 \Rightarrow \mathsf{i/j}$ after $[0,5]*$	0.39
$126.98 \leq \mathsf{wl(ijklm)} \leq 128.16 \Rightarrow \mathsf{j/i}$ after $[0,5]*$	0.36
$126.89 \leq \mathsf{wl(ijklm)} \leq 126.92 \Rightarrow \mathsf{j/k}$ after $[4,5]$	0.26
$124.67 \leq \mathsf{wl(np)} \leq 124.75 \Rightarrow \mathsf{n/i}$ after $[0,5]$	0.26
$1.60 \leq \mathsf{wl(or)} \leq 6.75 \Rightarrow \mathsf{o/n}$ after $[0,5]$	0.46
$3.02 \leq \mathsf{wl(or)} \leq 6.75 \Rightarrow \mathsf{r/o}$ after $[0,5]$	0.62
$2.24 \leq \mathsf{wl(stuv)} \leq 6.40 \Rightarrow \mathsf{s/r}$ after $[0,5]$	0.42
$2.33 \leq \mathsf{wl(stuv)} \leq 6.40 \Rightarrow \mathsf{u/s}$ after $[0,5]$	0.58
$2.78 \leq \mathsf{wl(stuv)} \leq 6.40 \Rightarrow \mathsf{v/u}$ after $[0,5]$	0.48

Table 5. The best rules discovered for each upstream movement effect. *Note that zones i and j meet at a confluence, so it is possible to move upstream into j from i and upstream into i from j

7 Conclusions and further work

We have developed the foundations of an algorithm that is able to identify the instances of rules of a particular logical form that best describe a given data-set. The approach can handle a range of logical forms, including simple causation, causation with conditional rules, and perpetuation. Our experiments show that for synthetic data, where the underlying causal rule is known, the approach is able to derive close approximations of the underlying causal rules from data.

In the case of real data, however, granularity effects may often confound an attempt to derive strict causation, where one event initiates another. In our example of fish movement, for example, the spatial and temporal granularity of the data (movement between granular zones of tens of kilometers and with a finest temporal granularity of one day), our algorithm struggles to identify strict causal relationships. However, by tasking the algorithm to look instead for perpetuation rules (termed in our system "Always-rules"), the algorithm is able to identify a suite of rules that compactly describe the data. Amongst our key results are included, considering rules relating fish movement to water level alone:

– the rules generated do relate to meaningful structure in the movement data, describing movements that are significantly different from random movements;
– the top-ranked rules in each zone compactly describe approximately 20% of the fish movements.

While this study has demonstrated the potential of our approach, future work on a much wider range of data sets is needed to further validate our initial results (in particular with finer-granularity information for events and process variables). Beyond this, comparison with probabilistic alternatives would assist both in validating our results and in further elucidating the practical implications of our logical approach. In the longer term, an integration of both probabilistic and logical approaches may be advantageous. It is also likely that, in moving towards operational data-mining tools for identifying causal relationships in movement data, we can complement our algorithm with visualisation capabilities for assisting users with sorting and filtering inferred rules. More broadly, we believe that visualisation of causal spatial rules could be a fruitful area for future research.

Finally, it is worth reflecting on the secondary role played in our account of causation by space, when compared with time. Cause and time are intimately linked through the familiar maxim that an effect cannot precede its cause, a reflection of the asymmetrical directedness of time. Since space exhibits no such directionality, there is no comparable maxim relating cause and space. Space and time do, however, share the attribute of extension, which gives rise to the measures of distance and duration. A general expectation for causality is that causal influence should be proximal with respect to both space and time: that is, we expect an effect to be spatially and temporally close to its cause (compare our remarks in [3] commenting on [6]). Where we find that this is apparently not the case — where a cause at one place and time leads to an effect at a distant place after a time delay — we normally suppose this to be explicable in terms of some unobserved process carrying the causal influence from the cause location to the effect location. But precisely because the process is unobserved, it is not possible for a mechanism that extracts causal rules from data to detect it, with the result that spatial linkages between cause and effect may, at least for some types of data set, show up only weakly, if at all, in the analysis.

Acknowledgments

Antony Galton's work was supported by the EPSRC, project EP/M012921/1. Matt Duckham's work is supported by funding from the Australian Research Council (ARC) under the Discovery Projects Scheme, project DP120100072. Alan Both's work is supported by funding from ARC project DP120103758.

References

1. E. Allen, G. Edwards, and Y. Bédard. Qualitative causal modeling in temporal GIS. In *Spatial Information Theory: A Theoretical Basis for GIS, Proceedings*

of COSIT'95, volume 988 of *Lecture Notes in Computer Science*, pages 397–412. Springer, 1995.
2. James F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.
3. S. Bleisch, M. Duckham, A. Galton, P. Laube, and J. Lyon. Mining candidate causal relationships in movement patterns. *International Journal of Geographical Information Science*, 28(2):363–382, 2013.
4. M. Bunge. *Causality*. Dover, New York, 1966.
5. Donald Davidson. Causal relations. *Journal of Philosophy*, 64:691–703, 1967.
6. B. A. El-Geresy, A. I. Abdelmoty, and C. B. Jones. Spatio-temporal geographic information systems: A causal perspective. In Y. Manolopoulos and P. Návrat, editors, *Proceedings of the 6th East European Conference on Advances in Databases and Information Systems (ADBIS)*, pages 191–203, Berlin, 2002. Springer.
7. A. Galton. States, process and events, and the ontology of causal relations. In M. Donnelly and G. Guizzardi, editors, *Formal Ontology in Information Systems: Proceedings of the Seventh International Conference (FOIS 2012)*, pages 279–292. IOS Press, 2012.
8. J. P. Lyon. Snags underpin Murray River restoration plan. *ECOS*, 177, 2012.
9. M. Moens and M. Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14:15–28, 1988.
10. J. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
11. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Springer-Verlag, New York, 1993.
12. Paolo Terenziani. Towards a causal ontology coping with the temporal constraints between causes and effects. *International Journal of Human-Computer Studies*, 43:847–863, 1995.
13. Paolo Terenziani and Pietro Torasso. Time, action-types, and causation: An integrated analysis. *Computational Intelligence*, 11(3):529–552, 1995.